Vendor: Databricks

Exam Code:DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST

Exam Name: Databricks Certified Professional Data Scientist Exam

Version: Demo

QUESTION 1

Of all the smokers in a particular district, 40% prefer brand A and 60% prefer brand B. Of those smokers who prefer brand A. 30% are females, and of those who prefer brand B. 40% are female. What is the probability that a randomly selected smoker prefers brand A, given that the person selected is a female?

Which of the following is a best way to solve this problem?

A. Bays Theorem

B. Poisson Distribution

C. Binomial Distribution

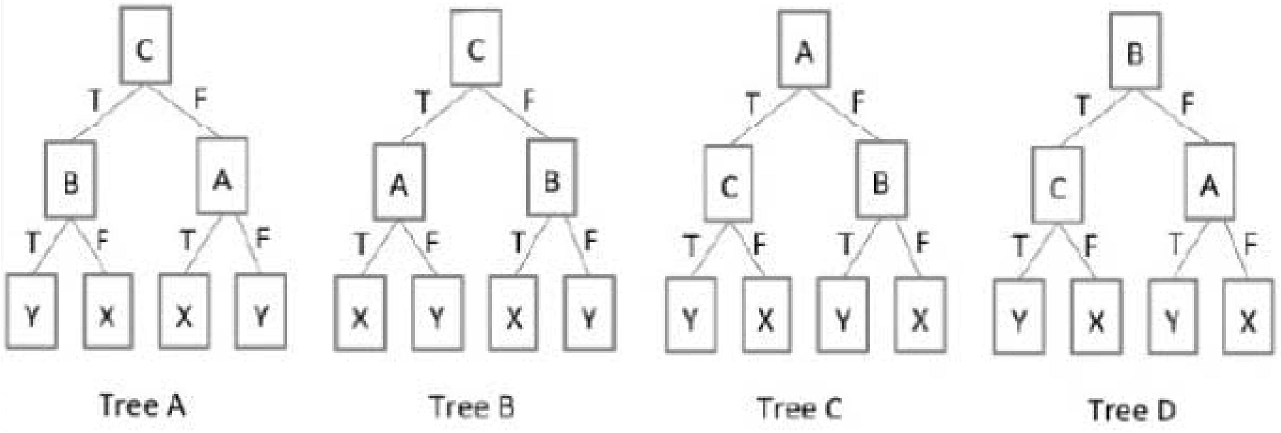
D. None of the above

Correct Answer: A

QUESTION 2

Refer to the Exhibit.

C	B	A	CLASS
T	T	T	X
T	T	F	Y
T	F	T	X
F	F	F	Y
F	T	T	X
F	F	T	Y



In the Exhibit, the table shows the values for the input Boolean attributes "A", "B", and "C". It also shows the values for the output attribute "class". Which decision tree is valid for the data?

A. Tree A

B. Tree B

C. Tree C

D. Tree D

Correct Answer: B

QUESTION 3

You are using k-means clustering to classify heart patients for a hospital. You have chosen Patient Sex, Height, Weight, Age and Income as measures and have used 3 clusters. When you create a pair-wise plot of the clusters, you notice that there is significant overlap between the clusters. What should you do?

A. Identify additional measures to add to the analysis

B. Remove one of the measures

C. Decrease the number of clusters

D. Increase the number of clusters

Correct Answer: C

QUESTION 4

You are creating a model for the recommending the book at Amazon.com, so which of the following recommender system you will use you don\\'t have cold start problem?

A. Naive Bayes classifier

B. Item-based collaborative filtering

C. User-based collaborative filtering

D. Content-based filtering

Correct Answer: D

Explanation: The cold start problem is most prevalent in recommender systems. Recommender systems form a specific type of information filtering (IF) technique that attempts to present information items (movies, music, books, news, images, web pages) that are likely of interest to the user. Typically, a recommender system compares the user\\'s profile to some reference characteristics. These characteristics may be from the information item (the content-based approach) or the user\\'s social environment (the collaborative filtering approach). In the content-based approach, the system must be capable of matching the characteristics of an item against relevant features in the user\\'s profile. In order to do this, it must first construct a sufficiently-detailed model of the user\\'s tastes and preferences through preference elicitation. This may be done either explicitly (by querying the user) or implicitly (by observing the user\\'s behaviour). In both cases, the cold start problem would imply that the user has to dedicate an amount of effort using the system in its \\'dumb\\' state - contributing to the construction of their user profile - before the system can start providing any intelligent recommendations. Content-based filtering recommender systems use information about items or users to make recommendations, rather than user preferences, so it will perform well with little user preference data. Item-based and user-based collaborative filtering makes predictions based on users\\' preferences for items, os they will typically perform poorly with little user preference data. Logistic regression is not recommender system technique.

QUESTION 5

Classification and regression are examples of___________.

A. supervised learning

B. un-supervised learning

C. Clustering

D. Density estimation

Correct Answer: A

Explanation: In classification, our job is to predict what class an instance of data should fall into. Another task in machine learning is regression. Regression is the prediction of a numeric value. Most people have probably seen an example of regression with a best-fit line drawn through some data points to generalize the data points. Classification and regression are examples of supervised learning. This set of problems is known as supervised because we\\'re telling the algorithm what to predict.

QUESTION 6

A website is opened 3 times by a user. What is the probability of he clicks 2 times the advertisement, is best calculated by"

A. Binomial

B. Poisson

C. Normal

D. Any of the above

Correct Answer: A

Explanation: In a binomial distribution, only 2 parameters, namely n and p, are needed to determine the probability. Where p is the probability of success and q is the probability of failure in a binomial trial, then the expected number of successes in n trials. This is a binomial distribution because there are only 2 possible outcomes (we get a 5 or we don\\'t).

QUESTION 7

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn\\'t rain, he incorrectly forecasts rain 10% of the time. Which of the following will you use to calculate the probability whether it will rain on the day of Marie\\'s wedding?

A. Naive Bayes

B. Logistic Regression

C. Random Decision Forests

D. All of the above

Correct Answer: A

Explanation: The sample space is defined by two mutually-exclusive events - it rains or it does not rain. Additionally, a third event occurs when the weatherman predicts rain. You should consider Bayes\\' theorem when the following conditions exist. ?The sample space is partitioned into a set of mutually exclusive events {A1, A2,... :An}. ?Within the sample space, there exists an event B: for which P(B)>; 0. ?The analytical goal is to compute a conditional probability of the form: P (Ak B).

QUESTION 8

You are working with the Clustering solution of the customer datasets. There are almost 40 variables are available for each customer and almost 1.00,0000 customer\\'s data is available. You want to reduce the number of variables for clustering, what would you do?

A. You will randomly reduce the number of variables

B. You will find the correlation among the variables and from their variables are not co- related will be discarded.

C. You will find the correlation among the variables and from the highly co-related variables, you will be considering only one or two variables from it.

D. You cannot discard any variable for creating clusters.

E. You can combine several variables in one variable

Correct Answer: CE

Explanation: When you are applying clustering technique and you find that there are quite a huge number of variables are available. Then it is better the find the co-relation among the variables and consider only one or two variables from the highly co-related variables. Because highly co-related variable will have the same effect, while creating the cluster. We can use scatter plot matrix among the variables to find the co-relation. You can also combine several variables into a single variable. For example if you have two values in the dataset like Asset and Debt than by combining these two values like Debt to Asset ratio and use it while creating the cluster.

QUESTION 9

What describes a true limitation of Logistic Regression method?

A. It does not handle redundant variables well.

B. It does not handle missing values well.

C. It does not handle correlated variables well.

D. It does not have explanatory values.

Correct Answer: B

QUESTION 10

What are the advantages of the mutual information over the Pearson correlation for text classification problems?

A. The mutual information has a meaningful test for statistical significance.

B. The mutual information can signal non-linear relationships between the dependent and independent variables.

C. The mutual information is easier to parallelize.

D. The mutual information doesn\\'t assume that the variables are normally distributed.

Correct Answer: C

Explanation: A linear scaling of the input variables (that may be caused by a change of units for the measurements) is sufficient to modify the PCA results. Feature selection methods that are sufficient for simple distributions of the patterns belonging to different classes can fail in classification tasks with complex decision boundaries. In addition, methods based on a linear dependence (like the correlation) cannot take care of arbitrary relations between the pattern coordinates and the different classes. On the contrary, the mutual information can measure arbitrary relations between variables and it does not depend on transformations acting on the different variables. This item concerns itself with feature selection for a text classification problem and references mutual information criteria. Mutual information is a bit more sophisticated than just selecting based on the simple correlation of two numbers because it can detect non- linear relationships that will not be identified by the correlation. Whenever possible: mutual information is a better feature selection technique than correlation. Mutual information is a quantification of the dependency between random variables. It is sometimes contrasted with linear correlation since mutual information captures nonlinear dependence. Correlation analysis provides a quantitative means of measuring the strength of a linear relationship between two vectors of data. Mutual information is essentially the measure of how much "knowledge" one can gain of a certain variable by knowing the value of another variable.

QUESTION 11

A. Naive Bayes classifier

B. Collaborative filtering

C. Logistic Regression

D. Content-based filtering

Correct Answer: B

Explanation: One scenario of collaborative filtering application is to recommend interesting or popular information as judged by the community. As a typical example, stories appear in the front page of Digg as they are "voted up" (rated positively) by the community. As the community becomes larger and more diverse, the promoted stories can better reflect the average interest of the community members.

QUESTION 12

You are working in an ecommerce organization, where you are designing and evaluating a recommender system, you need to select which of the following metric wilt always have the largest value?

A. Root Mean Square Error

B. Sum of Errors

C. Mean Absolute Error

D. Both land 2

E. Information is not good enough.

Correct Answer: E